

DOCUMENT RESUME

ED 126 102

TM 005 240

AUTHOR Weber, Margaret B.
TITLE Alternative Education: Implications for Evaluation Methodology.
PUB DATE [76]
NOTE 61p.; Paper presented at the Annual Meeting of the American Educational Research Association (60th, San Francisco, California, April 19-23, 1976); Not available in hard copy due to marginal legibility of original document

EDRS PRICE MF-\$0.83 Plus Postage. HC Not Available from EDRS.
DESCRIPTORS Achievement Gains; Affective Objectives; Affective Tests; Behavioral Objectives; Cognitive Objectives; Cognitive Tests; Comparative Analysis; *Curriculum Evaluation; Diagnostic Tests; *Educational Alternatives; *Evaluation Methods; Experimental Groups; Formative Evaluation; Hypothesis Testing; Individualized Programs; *Instructional Innovation; *Models; Post Testing; Pretesting; Program Effectiveness; Program Evaluation; Research Design; *Secondary Education; Statistical Analysis; Student Evaluation

IDENTIFIERS Community for Alternatives in Education; Illinois (Winnetka); New Trier West High School

ABSTRACT

An application of evaluation methodology to innovative educational programs through a case study of a Chicago suburban high school program for alternative education is presented. The evaluation methodology encompasses the difficulty in combining comparative programs, individual-diagnostic, and individual developmental measures, the selection of appropriate comparison groups and the definition of comparative program objectives, the psychometric issues of test selection, and the rationale for using a sophisticated statistical analysis to compensate for the absence of randomization. Reported data include standardized cognitive measures and both standardized and teacher-made affective measures.

(Author/DEP)

* Documents acquired by ERIC include many informal unpublished *
* materials not available from other sources. ERIC makes every effort *
* to obtain the best copy available. Nevertheless, items of marginal *
* reproducibility are often encountered and this affects the quality *
* of the microfiche and hardcopy reproductions ERIC makes available *
* via the ERIC Document Reproduction Service (EDRS). EDRS is not *
* responsible for the quality of the original document. Reproductions *
* supplied by EDRS are the best that can be made from the original. *

ED126102

BEST COPY AVAILABLE

TM005 240

ALTERNATIVE EDUCATION: IMPLICATIONS
FOR EVALUATION METHODOLOGY

by

Margaret B. Weber
Emory University

Paper Presented at the
American Educational Research Association
April 22, 1976

ACKNOWLEDGEMENTS

This study was done under unusual circumstances in a very unusual school system. The Community for Alternatives in Education had no provisions for evaluation and, as a result, was willing to permit a student of evaluation methodology to "experiment" with the program. In return for assistance throughout the planning and implementation phases (June 1974 through July 1975), the Community cooperated completely with this author in making all data available.

Many thanks to John Quimby Easton for his assistance in designing the study, analyzing the data, and discussing the outcomes. Finally, credit for any insightful parts of this paper should be shared with Susan S. Stodolsky who initially interested this author in evaluation and who guided this author's study of evaluation methodology during doctoral work at the University of Chicago.

March 10, 1976

M. B. Weber

The purposes of this paper are to present: (1) an evaluation model adapted from traditional curriculum evaluation models; (2) the application of this model to an innovative secondary program; and, (3) the evaluation model that derives from the results of this application. The goal of the evaluation presented here was never to defend or justify the value of the innovative program. The evaluation was, instead, undertaken to demonstrate that this "value" could be translated into values expressed by the extent public secondary program. The methodology was, in part, selected since it corresponded to evaluation efforts already incorporated into the school system which housed the innovative program and which would, therefore, be easily communicated. Had the innovative program faculty been less defensive and the evaluator more inventive, however, this paper might begin where it now ends.

Traditional Evaluation Models

Traditional evaluation methodology derives from Tyler's model for curriculum design. That model included definition of program purpose, specification of relevant instructional experiences, organization of these experiences, and evaluation of program objectives (Tyler, 1971). Tyler's model was intended for use in the traditional classroom.¹

In the decade of the sixties, instructional modernization fostered a liberalization of the traditional structured-curriculum for which Tyler's model had been designed. Individual learners were permitted to progress through a common set of educational objectives at their own rate.

This necessitated a modification of the Tyler model to permit evaluation of the individual learner in relation to a set of common objectives. Following is a modification presented by Lindvall and Cox (Tyler, 1969, 161):

1. Sequences of instructional objectives to define the curriculum.
2. Instructional materials to teach each objective.
3. An evaluation procedure for placing each pupil at the appropriate point in the curriculum.
4. A plan for developing individualized programs of study.
5. A procedure for evaluating and monitoring individual development.

This model elaborates on Tyler's model by adding a formative evaluation element and by redefining the Tyler model specifically in terms of the individual learner.

If the Lindvall and Cox model is implemented, judgments can be made about the relative "fit" between a student and a program. However, the question of "program effectiveness", without references to specific learners, cannot be answered by this model. Program effectiveness can only be judged with reference to an established criterion or with reference to an acceptable existing program.

The gains of the new program must be compared with the gains attributed to the criterion program. Thus, in the evaluation of an innovative program, one must ask not only is the students' progress through the program and at what rate but also would they progress more efficiently toward a stated goal if in another program? Attempts to answer this latter question push one beyond the Lindvall and Cox model.

3.

Following is an extension of the Lindvall and Cox model which incorporates summative comparative data. It was this model that was used to evaluate the innovative program discussed in the next section of this paper.

DESCRIPTIVE AND COMPARATIVE
EVALUATION MODEL

<u>Behaviors</u>	<u>Programs</u>		
	Comparison Program	Program Inter- section	"New" Program
Articulation of Educational Objectives			
Articulation of Behavioral Objectives	X		X
Identification of Common Behavioral Objectives	X		X
Development of Diagnostic Pre- and Post-Measures		X	
Development of Comparative Pre- and Post-Measures	X		X

Development of Instructional Materials	X		X
Formative Evaluation (*)	X		X
Instructional Revision (*)	X		X
Summative Evaluation	X	X	X

Note: Steps marked with asterisks (*) are repeated periodically throughout program implementation.

The process of evaluation in this model is complex. Not only is adequate measurement required so that each individual learner is appropriately described in terms of behavioral objectives, but comparative information is required in order to determine the differential effects of the program. Thus, one is faced with a three-fold task: (1) initial measurement of individuals prior to program onset for diagnostic purposes; (2) repeated measurement of individuals at specified times during the program to determine individual development; and (3) pre- and post-test measurement of the learners under study and of specified comparison groups in order to determine program "effectiveness."

Tasks #1 and #2 above as measurement issues are relatively easy to implement. Task #3, however, presents some obstacles to the evaluator. The identification of an appropriate comparison program is difficult. One is tempted to identify any existing program as a standard against which the "new" program is to be compared and this approach could possibly yield

relevant information about the relative merits of the new program.

However, with respect to programs which offer "new" types of learning experiences to the students, it is often difficult to find a meaningful comparison program.

The problem of comparison group identification is inherent in the evaluation of "new" programs. Attempts to change the process of education and thereby redefine the goals of education make identifying a comparison group sharing objectives difficult. It is more difficult to identify a comparison program that agrees with priorities for behavioral objectives shared with an innovative program. And, the more novel or innovative that the program under study is, the more difficult it becomes to find an appropriate comparison. This problem is exemplified in the case study reported here.

There is little question in the research community today that the 'experiment' provides the optimally sound basis for educational policy decisions. However, the feasibility of randomization, the sine qua non of experimental design, may be limited for local evaluation. In such instances the problem for the evaluator becomes one of making the best of a bad situation. Rather than testing hypotheses or estimating effects, the evaluator may have to fall back upon traditional evaluation methodology. . . (Weber, in press).

Clearly, to sample at random from a specified population and then to assign at random to treatment and control would be a delightful approach to local evaluation studies. However, what does one do when this is not possible? The following case study presents an alternative approach to

design which recognizes the limitations imposed by the absence of randomization, the limitations imposed by the small sample size, and which uses hypothesis testing and the estimation of effects to present a descriptive rather than inferential study.

Case Study: "Community for Alternatives in Education"

In attempting to evaluate an innovative program, one must design an optimal mix of program comparative, individual-diagnostic, and individual development or change measures. Because of the essentially divergent aspects of group comparative and individual-diagnostic specifications in a non-structured curriculum setting, it is particularly difficult to bring together individual evaluation which reflects individual development and curriculum evaluation which, in an innovative program, reflects the effectiveness of a new philosophy of learning or teaching.

This section of the paper will present the process chosen to provide curriculum evaluation information followed in a non-structured-curriculum setting. The particular setting described here is a Chicago suburban high school where a group of innovative and enthusiastic teachers have attempted to provide alternative patterns of learning to their students and to measure the impact of these patterns. By relating the process followed to select and construct instruments for the evaluation of this program, the deficiencies in traditional evaluation designs, as typified by the Tyler model and by the Lindvall and Cox modification as well as the limitations of the comparative model are illustrated.

Innovative Program

The "Community for Alternatives in Education" derives from a philosophy of learning which provides a student-structured learning environment. Student behavior is self-directed, content areas are self-selected, and teachers function not as instructors but as facilitators of learning experiences. This is contrasted to the more traditional learning environment typified by the "parent" school, i.e. the school in which the Community is housed. The parent school places its emphasis on cognitive development in an individualized environment. The Community encourages cognitive and affective learning experiences whereas the parent school generally encourages only cognitive outcomes.

Following are the educational objectives formulated by the Community for Alternatives in Education:

1. To form a learning community that offers students alternatives in education different from those offered by the parent school.
2. To create, as a group, a structure which allows each individual to choose his/her own learning style, but which, at the same time, brings to him/her the feeling of belonging, and the affective ties of a small community.
3. To involve the individual in decisions concerning his/her learning needs and the process by which these needs are to be met.
4. To make the individual responsible for achieving his identified learning needs.

5. To encourage self-evaluation.
6. To make the processes of learning and the development of the total person as important as the knowledge acquired.
7. To create a learning situation in which the individual can follow his objectives across subject area divisions, unconstrained by time and space.

These seven educational objectives provide insight to the learning environment fostered by the Community and to the behaviors anticipated in that environment. The seven program objectives were reviewed and translated into six evaluation areas. Following is the list of evaluation areas derived from this review:

1. Evaluation of the learner with respect to stated objectives including self-evaluation and the evaluation of the Community advisors of student performance.
2. Diagnostic evaluation of the learners on specific academic skills (reading, math, etc.).
3. Evaluation of the Community as a whole and of a comparison group(s) with respect to specific academic skills (reading, math, etc.).
4. Evaluation of the Community as a whole and of a comparison group(s) with respect to specific affective objectives.
5. Evaluation of the Community for tolerance of diverse learning styles.²
6. Evaluation of the Community for non-specified outcomes.

Once this focus had been established, five of the six areas listed above were reformulated into behavioral objectives by first specifying cognitive behaviors of interest to the Community and to the parent school which would provide the comparison group(s), and by specifying affective behaviors of interest to the Community. These behaviors included the following:

Cognitive: Reading, English Usage,
Numerical Competency,
Critical Thinking,³ and
Creativity.

Affective: Attitudes toward Teachers,
Attitudes toward Peers,
Attitudes toward Competition,
Attitudes toward Learning,
Locus of Control, and Tolerance
for Learning Styles.

The first evaluation area listed above was not included in the evaluation design. The Community faculty felt that the individualistic nature of the content and behavior of student's learning objectives necessitated individualistic evaluation procedures. These procedures were to be determined by each student in cooperation with her/his Community advisors.

The cognitive and affective behaviors were translated into the following behavioral objectives:

1. Students participating in the Community will recognize teachers as individuals who help students define, reach, and evaluate their goals.
2. Students participating in the Community will evidence a decline in anxiety expressed about evaluation procedures and begin to redefine evaluation as a learning aid.

3. The notion of competition will transfer from that of intraindividual to interindividual.
4. Peer relationships will be based on shared interests without regard to achievement level or aptitude.
5. The student group will evidence a wide variety of learning styles in response to individual preferences.
6. Students will evidence an increase in positive attitudes toward "school" as an institution.
7. The number of interests and/or depth of interests will increase within individual students and across the student group.
8. The types of interests pursued by the student group will be different from those interests pursued in the parent school.
9. Students will evidence an increase in internal locus of control, i.e. sense of control over their own fate.
10. Students from the Community will evidence behavior similar to that of the comparison group(s) on measures of traditional academic skills over the period of one academic year.

Suitable published instruments were available for Objective #10 that could be adapted to yield individual diagnostic and developmental information as well as group information for comparative purposes. Buros' Mental Measurement Yearbook (1972) was used to formulate a general reference list for such cognitive skill tests appropriate for high school students. This list included tests of general reading, English, and mathematics achievement as well as tests of critical thinking and problem solving.⁴ The participating Community teachers reviewed test specimens.

The final selection was based on this review.

A similar list from Buros' Mental Measurement Yearbook (1972) was compiled for the nine affective objectives. An appropriate instrument was identified for Objective #9. However, due to the scarcity of well-developed appropriate instruments for the other affective objectives, teacher-made tests were developed specifically for the objectives of the Community.

The construction of teacher-made affective instruments resulted in a less sophisticated measurement process for the majority of affective objectives than that permitted for the cognitive objective. This resulted from a lack of opportunity to pilot the teacher-made tests prior to their inclusion in this evaluation.

For each of the eight affective objectives measured in this way, a redefinition of the characteristic was provided in terms of the item content of specific tests. No assumptions were made that the items were sampled from a universe of items and, because no test pilots were possible, neither estimates of reliability or validity of the tests beyond simple estimates of content validity were available.

The selection or construction instruments for this evaluation was influenced by a desire for economy, student interest, and format variety. Three instruments were developed. Each had a different format, and each contained items from more than one affective area as defined by the Objectives #1 through #8. A quick view of the teacher-made tests will reveal the combination of objectives measured by them. The three teacher-made instruments included a Likert-type scale, an adjective checklist, and

an open-ended questionnaire. The Rotter Internal/External Locus of Control was used to address Objective #9. This instrument was designed to measure the extent to which an individual believes that he/she has control over his/her fate.

A fifth affective instrument was added to the post-test battery as a result of a review of the student questionnaires: Feedback from the Community students indicated that approximately 60% felt that highly structured learning environments, i.e. those classrooms or schools with specified times for specified content material, etc., interfered with the amount of learning that took place. Therefore, Ginther's shortened version of the "Inventory of Beliefs" (1973) was adopted for use in this study. It was expected that students who had elected to participate in the Community would respond differently to Ginther's instrument than would students who had not elected to participate, i.e. those students in the comparison groups would have remained in the parent school.

Evaluation

The objectives of the "new" program become important in the evaluation design as one attempts to make decisions about the relative "fit" between participating learners and the program as implemented. This fit would be periodically determined as part of the formative evaluation procedures. Following Lindvall and Cox (Tyler, 1971), a formative evaluation was planned using observational procedures. Psychometric measures were not planned for the formative evaluations because of time constraints and the inavailability of more than two parallel forms of the cognitive tests and none for the teacher-made or Rotter instruments. The success or failure of the program for individual learners is not

under discussion here, however. This paper presents the comparative data and interpretation only. Thus, only the ninth and tenth objective will be discussed.

The requirement for comparative data in program evaluation necessitates the formulation of a set of behavior-content specifications as suggested by Tyler (1971, 47) derived from the intersection of new and comparison program goals. By comparing post-test performance, having controlled for initial differences statistically, one has a legitimate basis for comparative judgment with respect to specified objectives. However, the intersection of new and comparison program goals or objectives need not reflect an intersection of priorities. And, with respect to this case study, the intersection does not reflect an intersection of priorities. This is an important deficit of a comparative design. For example, significant differences reported between programs may suggest that one has "failed" where another has "succeeded," the reality may be that the objective being compared was of low priority for the "failure" and of high priority for the "successful." Thus, significant statistical differences may demonstrate the validity of program priorities. This deficit in comparative program evaluation will be re-emphasized in the data analysis section.

Adequate evaluation of new programs requires the inclusion of objectives with a common base in new and comparison programs and objectives specific to the new program. Behavior-content specification is not adequate for evaluation of innovative programs. Instead, the requirement becomes a set of behaviors derived from the educational objectives from both new and comparison programs and a set of non-specific content area objectives

broad enough to be reflected in each individual self-directed learning experience. For example, the emphasis in the non-structured curriculum reported here is on generalizing behaviors such as organization, decision-making, inquiry skills, scientific methodology and creativity, etc. This derives from the philosophy of the Community. Skills based on specific content areas are not emphasized.

An evaluation design was selected that would yield post-measures comparisons on the ninth and tenth objective. The other objectives were to be evaluated through non-traditional measurement practices.

Following Scriven's goal-free evaluation scheme and incorporating it into this evaluation resulted in a combination of tests, questionnaires,⁵ and observation schedules designed for the Community for Alternatives in Education. Briefly, Scriven's goal-free method refers to program evaluation by observers who have no idea of program objectives. They evaluate the program only with respect to observed outcomes. Ultimately, it is the combination of structured and non-structured evaluation procedures that will tell educators the extent of the impact of innovative curriculum programs. The future of program evaluation in education depends on such precise and systematic attempts to measure group processes, individual skill development and environmental changes that theoretically accompany new curricula.

Description of Program and Comparison Groups

One-hundred and fourteen students enrolled in the Community for the 1974-75 academic year. Of these, eighty were seniors and thirty-four

were juniors. These students represented each of the four ability levels in the high school and these participating students expressed a wide range of interests they hoped to pursue in the Community. The commonality within this group was an expressed dissatisfaction with traditional education, i.e. the type of learning experiences offered in the parent school. All 114 Community students participated in the initial testing sessions. However, in both January and May a considerable number refused to take the re-tests. As a result, a total of fifty-nine students had complete data on pre- and post-tests.

The selection or identification of comparison groups was undertaken with no incentive for participation. In addition, because student participation in the Community was based on voluntary application, it was felt that a random sample drawn from the parent high school would provide no more meaningful a comparison than would comparisons made with groups of students recruited from the parent high school. Also, it was felt that post-test participation might be difficult to guarantee given a random sample. Therefore, a group of approximately one hundred juniors and seniors, not participating in the Community but sharing similar characteristics such as ability level, were invited to participate in this evaluation. Of these one hundred who were initially contacted, fifty-four agreed to participate. The breakdown of both Community and Comparison groups by sex, grade level, and group is available in Appendix A.

Selection of Post-test Instruments

Approximately eight hours of testing were required for the entire pre-test battery selected for this evaluation study. This battery included:

(1) Stanford Test of Academic Skills which includes two reading comprehension tests, five English usage tests and one mathematics test; (2) Cornell Test of Critical Thinking; (3) Tennessee Self-Concept Test; (4) Rotter Internal/External Locus of Control Test; (5) Teacher-made Likert-type Scale; (6) Teacher-made Adjective Ranking Test; and (7) Teacher-made Student Questionnaire. (See Appendix B for the teacher made tests.)

The pre-test results were item analyzed and a Kuder-Richardson₂₀ reliability estimate was derived for each of the first two tests listed above in order to select from the initial battery the most reliable and the most appropriate in terms of difficulty for a three-hour post-test battery. The Student Questionnaire was automatically included in the post-test battery without reliability data. (See Appendix C for KR₂₀ estimates.)

The third, fourth and fifth tests listed above were not item analyzed, but a Split-Half reliability estimate for each was derived (See Appendix D). The sixth test listed above was eliminated from the pre-test analysis because the majority of students failed to complete the test or failed to complete it correctly, i.e. failed to follow directions.

The post-test battery was designed by selecting only those instruments from the pre-test battery which had a reliability which was greater than or equal to .80. This reliability coefficient level was selected following Cronbach (1970). Thus, the following tests were included on the post-test battery: (1) Stanford Test of Academic Skills; (2) Cornell Test of Critical Thinking; (3) Tennessee Self-Concept; (4) Rotter Internal/External Locus of Control Test; and (5) Student Questionnaire. However, this battery exceeded the three hour time allotment for the post-test administration. As a result, correlations were calculated between the Stanford

reading and English tests (See Appendix E). These correlations indicated that the administration of either the reading section or the English section would be sufficient to represent verbal skill development. The final decision was to include the reading tests and to eliminate the English section from the post-test battery. This decision was based on recent research reported by Bloom (1976) suggesting that reading comprehension skills are critical to the development of other academic areas.

In addition to those tests mentioned above, a sixth test, not given in the pre-test battery, was included in the post-test administration. That test is Ginther's shortened version of the "Inventory of Beliefs" (1973). This test did not come to the attention of the program evaluator early enough to be included in the pre-test session.

Design

Since three distinct groups were expected to develop from the Community student body, three groups from the parent school were identified from the recruited comparison group.⁵ In addition, comparisons were planned between groups within the Community population. Although rigorous research practices suggest that six contrasts with the number of replicates available for this study are of little utility, the emphasis in this evaluation is on description and not inference. As such, it was felt that something might be gained by analyzing the data in this manner. The six groups are as follows:

- I. September to June Community Participants.
- II. September to January Community Participants leaving to graduate.⁶

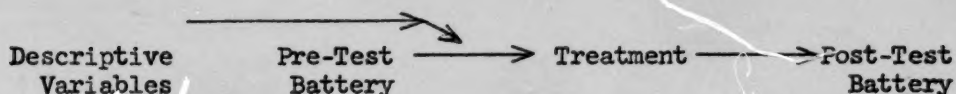
- III. September to January Community Participants leaving to return to the regular school.
- IV. September to June Parent School Participants.
- V. September to January Parent School Participants leaving to graduate.
- VI. September to January Parent School Participants leaving to return to the regular school.

Because the students from New Trier West High School were not randomly assigned to treatment groups, each group is biased. However, by using New Trier West students for the comparison groups, the six groups were assumed to be homogeneous with respect to socio-economic status.

In order to reduce bias, resulting from non-random assignment, covariate adjustments were planned for pre-test performance on the instruments included in the post-test battery (e.g. Campbell and Erlichson, 1973). In a study such as this, one is frequently faced with either reducing the bias with statistical methods or of retaining it. However, it is important to recognize that the adjustments may be increasing the error of estimate. The limitations of this design must be kept firmly in mind when using the results of this study for policy decisions.

Conceptual Model

Following is a schematic presentation of the conceptual model:



(See Table 1 for a complete listing variables)

Insert Table 1
about here

The six independent measures (including all Pre-Test Battery measures except the Student Questionnaire and the Descriptive variable "Grade Point Average") and the six dependent measures (including those measures listed under Post-Test Battery except for the Student Questionnaire) are continuous variables and are treated as such in the statistical analysis. The Descriptive Variables "Sex," "Grade Level," and "Treatment" are, however, treated as nominal variables and are used as factors in the design. The Student Questionnaire also is treated as nominal and/or ordinal and is analyzed in a separate qualitative section in a supplementary paper.

The exclusion of the Tennessee Self-Concept measure from the Pre- and Post-Test Battery reflects a practical psychometric judgment. Because only eighty-six complete data cases were obtained the analysis was already error-full and biased and, since the Tennessee Self-Concept measure is composed of fifteen separate subscales which suggest separate analysis, the evaluator decided to eliminate this test (e.g. Bock, 1975).

Statistical Analysis and Hypotheses

As this evaluation was not based on prior effort or experience, there was no empirical evidence that sex differences or cohort differences, i.e. grade level differences, would affect the outcomes. However, in order to identify such main effects and possible interactions, a three-way analysis of variance design was used. In addition, this analysis was multivariate in nature, as more than one outcome, i.e. dependent variable, was tested.

Table 1

Conceptual Model

Pre-Test Battery

Reading Part A
Reading Part B
Mathematics
Cornell Test of Critical Thinking
Rotter Internal/External Locus of Control
Student Questionnaire

Post-Test Battery

Reading Part A
Reading Part B
Mathematics
Cornell Test of Critical Thinking
Rotter Internal/External Locus of Control
Student Questionnaire
Short version of the Inventory of Beliefs

Descriptive Variables

Sex
Grade Level
Treatment Condition

Causal inference given the number of replicates per cell is questionable. However, the analysis does give one some insight into treatment effects and, the design itself has potential in local evaluation studies if a large N is available. This study also emphasizes the utility of multivariate analysis in local evaluation efforts. Thus, a three-way multivariate analysis of covariance (Grade Level X Sex X Group) was used to test the following hypotheses:

- H1: There are no significant differences among the six groups on the achievement measures: Reading Part A, Reading Part B, and Mathematics.
- H2: Groups #1, #2, and #3 have significantly higher mean scores on the Cornell Test of Critical Thinking than will groups #4, #5, and #6.
- H3: Groups #1, #2, and #3 have significantly higher mean scores on all affective measures than will Groups #4, #5, and #6.
- H4: There are no significant differences on any dependent variables between sex groups.
- H5: There are no significant differences on any dependent variable between grade levels.

The following hypotheses were tested using raw regression coefficients.

- H6: There is a significant positive relationship Grade Point Average and the Pre-Test score on the Cornell Test of Critical Thinking.

H7: There is a significant negative relationship between Grade Point Average and Post-Test scores on the Ratter Internal/External Locus of Control Test.

The advantages of covariate adjustments in testing the above hypotheses is that adjusted outcome measures, i.e. the dependent variables listed in the Post-Test Battery, can be compared. In other words, the statistical adjustment that covariate analysis provides equalizes the unequal entry characteristics of the subjects.

Although the multivariate analysis of covariance does help the evaluator "make the best of a bad situation," it does not permit descriptions of group "gains." Gain score analysis, which would answer questions mentioned in the first section of the paper on individual development is not appropriate unless random assignment has been used. Therefore, we are only able to ask comparative questions and thus to phrase only comparative hypotheses in this analysis. In short, we are only able to determine which group does better on the post-test measures, given adjustments to equalize pre-test differences.

Results

H1 was formulated to determine whether students participating in the innovative program evidenced less strong performance in areas of basic skills than did those students in the traditional program. The raw and adjusted means for the six groups on all outcome measures are reported in Table 2.

Insert Table 2
about here

As reported in Table 2, the multivariate F indicates that significant differences do exist among the groups on at least one outcome measure. An examination of the Step Down F values, however, indicates that there are no significant differences with respect to the three measures of basic skills. Thus, H1 is accepted.

The acceptance of H1 enables the Community to demonstrate that whatever the content of their program, clearly defined or poorly articulated, over the period of one academic year the participating learners are not detrimentally affected in reading or mathematics skill areas as operationalized by the Stanford Achievement battery. It is, however, interesting to note that basic skill acquisition, a high priority objective in traditional programs, did not elicit higher mean response level in groups 1, 2, or 3. What remains to be answered, however, is whether detrimental effects would occur over longer periods of time or with younger participants.

Also, as indicated in Table 2, the Cornell Test of Critical Thinking failed to discriminate among the six groups. Thus, H2 is rejected. The testing and resulting rejection of this hypothesis begins to explore the high priority objectives for the innovative program. Table 3 does indicate, that the two affective measures (Preference for Structure and the Rotter Internal/External Locus of Control Tests) did discriminate significantly ($p < .05$) than the scores of Groups I, II, IV, or V. This can be interpreted to mean that, all things being equal, the environment of Group VI, i.e. the environment

Table 2
Raw and Adjusted Means

Group	Variable	Raw Means	Adjusted Means
I	Reading Pt. A	38.92	38.19
I	Reading Pt. B	19.16	19.33
I	Mathematics	7.22	7.16
I	Rotter	3.67	4.61
I	Cornell	46.12	47.16
I	Preference for Structure	11.51	11.37
II	Reading Pt. A	43.67	40.59
II	Reading Pt. B	25.33	24.03
II	Mathematics	10.67	9.91
II	Rotter	6.00	6.05
II	Cornell	58.33	56.42
II	Preference for Structure	.00	.69
III	Reading Pt. A	42.40	42.93
III	Reading Pt. B	22.60	22.54
III	Mathematics	12.20	11.84
III	Rotter	4.72	4.72
III	Cornell	49.47	49.47
III	Preference for Structure	.00	-.77

25.

IV	Reading Pt. A	39.78	38.91
IV	Reading Pt. B	20.94	20.22
IV	Mathematics	6.67	5.91
IV	Rotter	4.61	4.31
IV	Cornell	43.17	42.98
IV	Preference for Structure	12.39	11.49
V	Reading Pt. A	41.50	39.73
V	Reading Pt. B	22.00	20.08
V	Mathematics	10.67	10.66
V	Rotter	6.17	3.48
V	Cornell	46.50	48.15
V	Preference for Structure	.00	.02
VI	Reading Pt. A	41.67	41.44
VI	Reading Pt. B	21.67	21.46
VI	Mathematics	10.67	11.16
VI	Rotter	.00	7.28
VI	Cornell	51.00	47.94
VI	Preference for Structure	.00	.52

Table 2a

The Effect of Group on the Six
Specified Outcomes

Multivariate $F = 3.7424$ $p < .0001$ $DF = 30 \text{ and } 246.000$

Source	Mean Square	Univariate F	p	Step Down F	p
Structure	332.9102	18.0748	.0001	18.0748	.0001*
Reading Pt. B	20.1548	1.8151	.1220	1.1163	.3606
Reading Pt. A	19.2317	.9688	.4435	.7221	.6094
Mathematics	48.7934	1.7296	.1401	.1577	.9769
Cornell	104.2604	1.4560	.2163	1.9786	.0943
Rotter	6.4869	2.0741	.0797	5.0007	.0007*

of the comparison group in the parent school fosters more of a sense that the individual is in control of his/her own fate than does the environment of the Community. Of course, because of the relatively small sample involved in this evaluation, it is possible that other factors entered into these results. However, one can conclude that the Community for Alternatives in Education is not effective in increasing internal Locus of Control as measured by the Rotter instrument. Again, the question of trait stability should be examined here as it is possible that sample bias coupled with trait stability accounts for these results.

 Insert Figure 1
 about here

The graph of the Preference for Structure Test (See Figure 2) also indicates that significant differences ($p < .05$) exist between groups. However, these differences are between groups categorized as September to January Community or comparison and all other groups.

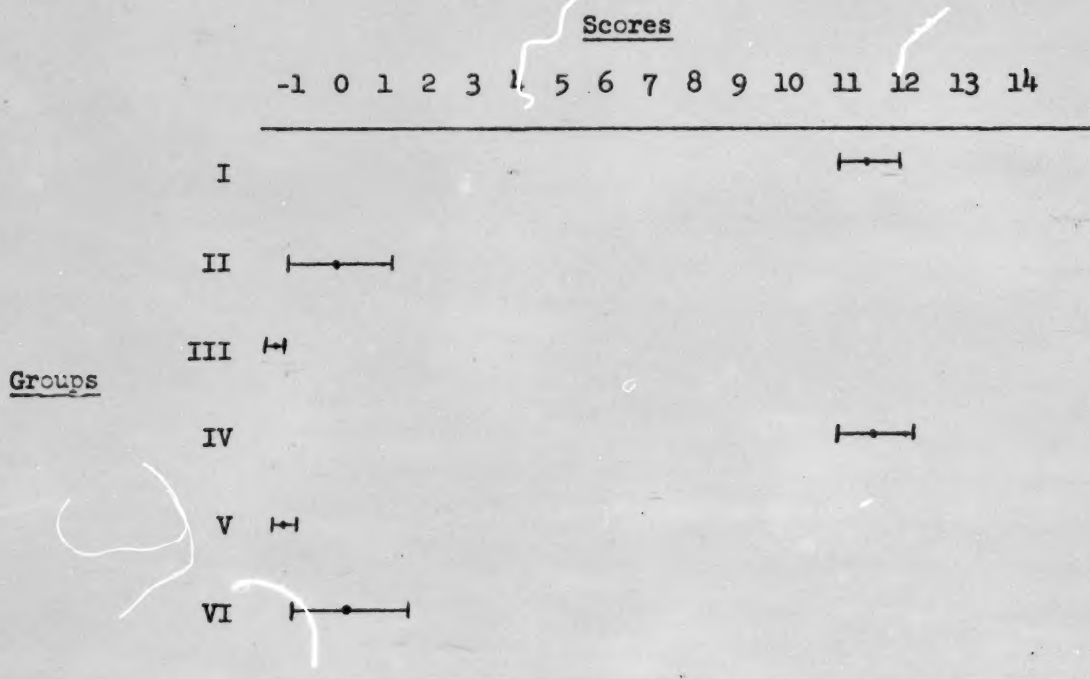
 Insert Figure 2
 about here

What we find is that both Community and comparison groups (I and IV) have higher scores or prefer less structure than do the other four groups. Since this test does not discriminate between groups in the Community and the groups in the parent school, H_3 is rejected.

In the multivariate analysis testing where "Sex" as a main effect was significant, no significant differences were found on any dependent variable. Therefore, H_4 was accepted.

Figure 1

Estimated Means for the Six Groups on the Preference
for Structure Test



However, a Sex X Group interaction was found (See Table 3) on the Rotter Internal/External Locus of Control Test. (See Table 4 for adjusted means). The graph of the means for both sexes and for all groups (See Figure 3) indicates that the girls in group V did better than the boys. In order to understand this interaction, further study of the students in group V would be necessary. And, as there are only seven replicates, the effect may be a function of sample bias.

Insert Table 3
about here

Insert Table 4
about here

Insert Figure 3
about here

Figure 2

Estimated Means for the Six Groups on the Rotter

Internal/External Locus of Control

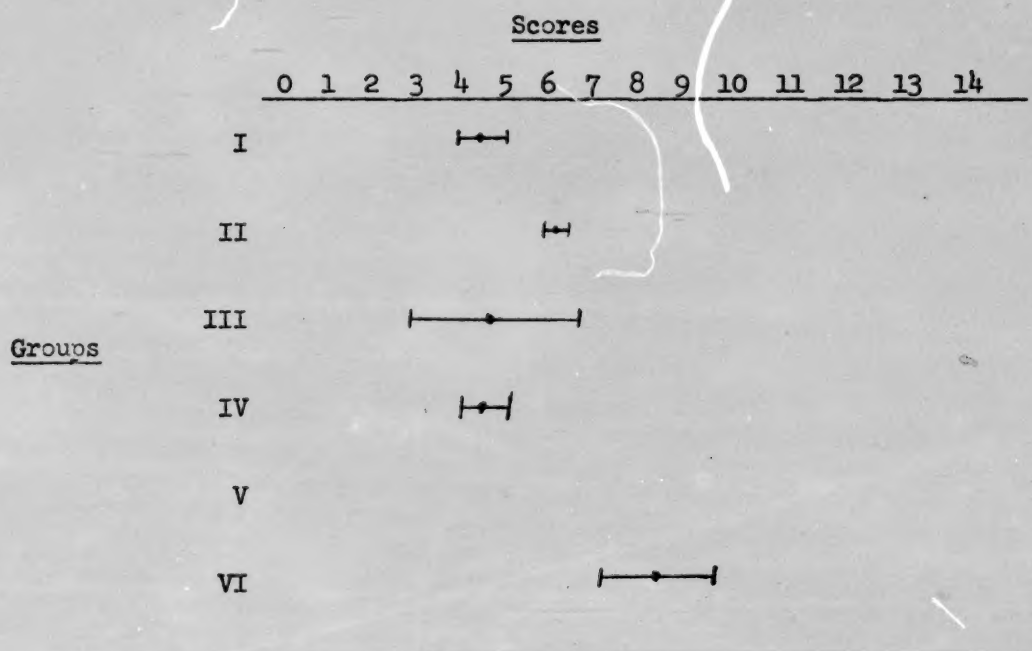


Table 3

The Effect of the Sex X Group Interaction
on the Specified Six Outcomes

Multivariate $F = 1.8122$ $p < .0222$ $DF = 18$ and 390.8081

Source	Mean Square	Univariate F	$p <$	Step Down F	$p <$
Structure	16.4932	.4926	.6881	.4926	.6881
Reading Pt. B	47.9049	.5239	.6666	.4843	.6937
Reading Pt. A	61.3343	.2548	.8579	.7479	.5254
Mathematics	14.2230	.5652	.6389	.6256	.5997
Cornell	205.2326	.4327	.7300	1.2289	.3016
Rotter	13.5393	2.3639	.0737	7.4988	.0002

Table 4

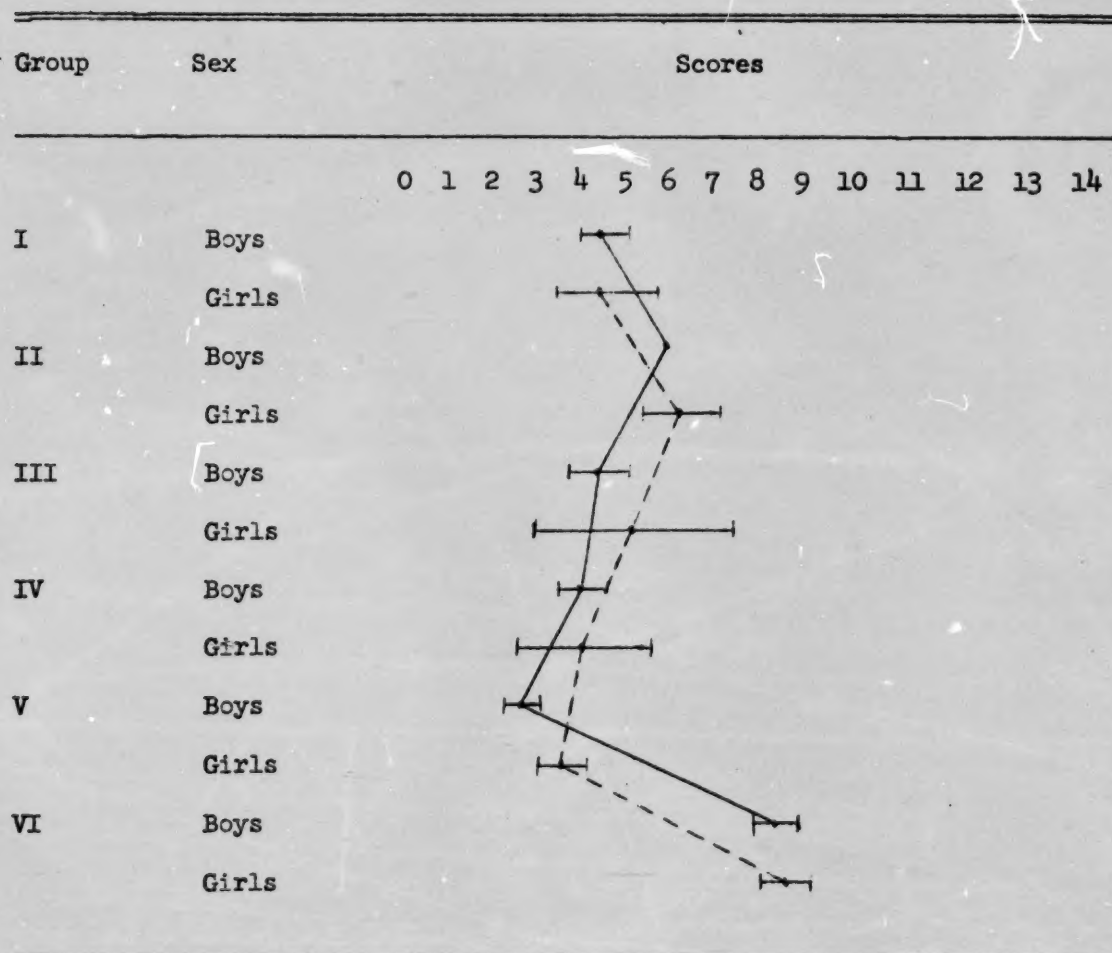
Raw and Adjusted Means for the Rotter Locus
of Control by Sex and Group

Group	Sex	Raw Mean	Adjusted Mean	Standard Error
I	Boys	4.37	4.501	.03
I	Girls	4.69	4.724	.29
II	Boys	*NA	5.934	.00
II	Girls	6.00	6.157	.19
III	Boys	6.50	4.611	.03
III	Girls	3.57	4.834	.68
IV	Boys	4.20	4.251	.55
IV	Girls	4.77	4.474	.03
V	Boys	*NA	3.372	.00
V	Girls	3.17	3.595	.03
VI	Boys	7.00	7.173	.24
VI	Girls	8.00	7.396	.03

*NA refers to an empty cell.

Figure 3

Graph of the Rotter Scores
by Group and Sex



By referring to Table 5, one finds that a significant positive relationship does exist between Pre-Test reading, as measured by Reading Part A, and Post-Test Critical Thinking scores. And, one would expect that reading ability is a necessary, although perhaps not sufficient skill for critical thinking behavior. This is particularly so given the specific instrument used to measure critical thinking in this study. The Cornell Test of Critical Thinking appears to be somewhat difficult to read.

 Insert Table 5
 about here

In addition, one finds that Pre-Test scores on Reading Part A are significantly and positively related to Reading Part B. Likewise, the Pre-Test score on the Cornell Test of Critical Thinking significantly predicts Post-Test performance on that same instrument.

Thus, in reviewing the hypotheses, we find that, although not inhibiting traditional skill acquisition, the Community for Alternatives in Education has not produced its intended effects. Several alternative explanations are possible: (1) the statistical analysis was not sufficient to eliminate or control for essential differences between the groups that tended to minimize the program's effects; (2) the instruments used in the Post-Test analysis were not valid in that they did not measure the phenomenon of the Community; (3) the number of students for whom complete data was available was not large enough to produce results favorable for the Community; or, the testing conditions were such as to produce invalid results. Certainly all of these

Table 5

Raw Regression Coefficients for Pre- and
Post- Measures
(N=114)

Pre-	Post-					
	Struc- ture	Read. Pt. A	Read. Pt. B	Math.	Cornell	Rotter
Grade Point Average	-.67 se=.46	-.17 se=.48	.09 se=.36	.12 se=.57	.39 se=.90	.05 se=.19
Read. Pt. A	.06 se=.10	.42* se=.10	.26* se=.07	.06 se=.12	.43* se=.19	.07 se=.04
Read. Pt. B	.36 se=.16	.23 se=.16	.13 se=.12	.11 se=.20	-.18 se=.31	-.01 se=.06
Math.	-.04 se=.07	.08 se=.07	.12 se=.06	-.12 se=.09	.06 se=.14	-.02 se=.03
Cornell	-.02 se=.07	.16 se=.08	-.08 se=.07	.03 se=.11	.43* se=.18	.00 se=.04
Rotter	.00 se=.09	.17 se=.09	-.03 se=.07	.10 se=.11	-.20 se=.18	.01 se=.04

* significant at p .05

alternative explanations are plausible and perhaps all of them produced what appears to be a failure of the Community to produce its intended outcomes.

Each of the above explanations suggests methods for improving future evaluation attempts, however. And, as this was only an exploratory evaluation, all is not lost. By examining the inter-test correlation matrix of the Post-Test scores with the independent variables eliminated or controlled for, some interesting relationships appear.

By referring to Appendix F. we find two obvious relationships: Both Reading Part A and Critical Thinking are positively and significantly associated with Reading Part B. However, there are also four other less obvious but completely logical relationships that may prove to be useful in selecting students who will benefit from the Community. These four relationships are: (1) A significant positive association between the Rotter and the Preference for Structure Test; (2) A significant positive association between the Post-Test Critical Thinking Test and the Preference for Structure; (3) A significant positive association between the Post-Test Rotter and the Post-Test Critical Thinking Test; and, (4) A significant negative association between the Post-Test Mathematics Test and the Preference for Structure Test.

The relationship between the Rotter and Preference for Structure indicates that individuals who have a large amount of Internal Locus of Control, i.e. who believe that they have control over their own fate, also have a strong preference for lack of structure.⁷ This is consistent with the expectation that people who feel in control of their

lives would be more likely to effectively function in and benefit from unstructured environments. Indeed, these people would probably find it very difficult to function effectively in an environment which is highly structured. The relationship between Critical Thinking and Preference for Structure can be accounted for in a fashion similar to that used to account for the relationship between the Rotter and Preference for Structure. Again, one would expect that individuals who are able to synthesize logical arguments from large amounts of data would find a highly structured environment somewhat constrictive. They would, however, benefit from an unstructured situation.

The positive relationship between the Rotter and Critical Thinking indicates that individuals who are analytical feel in control of their lives. Of course, since this relationship is only one of association, further study would be required to determine which characteristics produce or make possible the other.

The fourth relationship mentioned is that existing between Mathematics achievement and Preference for Structure. This significant negative relationship suggests that individuals who do well on mathematics achievement tests prefer to live and work in a highly structured environment.

Based on these four measures of association, subsequent evaluation efforts might profitably include using the Rotter and the Preference for Structure tests as screening devices if critical thinking is retained as a program objective. Similarly, one might suggest that students interested in the study of mathematics be screened from program participation unless some highly structured environment could

be implemented for those students.

Conclusion

The weaknesses in this evaluation study are a product of practical decisions. The major flaw is the inability to design an experimental study. The constraints resulting from this inability reduce the interpretations that can be drawn from the statistical analysis. And, the important study of gain scores was prevented as a result of the quasi-experimental design. Other major constraints, however, from outside the design itself were present. For example, neither Community participants nor the comparison group participants appeared to take the testing seriously. This results in a sizable number of incomplete data cases.

There are more important areas of weakness, however, that lie within the structure of the Community for Alternatives itself. As in any innovative, untested, and experimental program, the people involved in defining the Community objectives and of structuring the learning environment were left to their own devices. No "experts" from similar programs were made available to the Community faculty during the planning or operation of the Community's first year. Neither was technical assistance for psychometric analyses, etc., available. It seems particularly ironic to facilitate the development of a program of alternative education without facilitating the means for measuring its effectiveness. The evaluation study of the Community of Alternatives

in Education has merit despite many limitations. First, the faculty involved felt it worthwhile to objectively summarize their year's efforts and to communicate this summary to others. Second, the faculty learned a great deal about the identification, selection and interpretation of various psychometric methods. Third, objectives were clarified, measured, and redefined. And fourth, ideas have arisen for better evaluation practices for the coming year. (See Appendix G for a list of suggestions for future evaluation efforts). However, perhaps the most important contribution to the area of evaluation methodology that resulted from the New Trier case study was a new evaluation model for innovative secondary programs.

An Innovative Evaluation Model

As has been pointed out in critiques of preschool intervention evaluation methodologies (e.g. Stodolosy, 1973), common problems in the evaluation of innovative individualized programs result. It appears that the outcomes of innovative individualized programs are dependent upon: (1) precise articulation of program objectives; (2) identification or development of valid instruments, and (3) control over learner experiences.

In reference to preschool intervention evaluation, (Stodolsky, 1973) suggests the following:

- (1) The more homogeneous the treatment,
the more likely effects will be formed,
- (2) the more heterogeneous the treatment,
the more likely effects will not
emerge for the group as a whole,

- (3) the more control over individual experiences, the more likely it is that outcomes can be related to articulated treatments.

There is no reason to suggest that an innovative secondary program would be less susceptible to these three statements than are innovative preschool programs. If this assumption is true, three methodological issues with respect to evaluation arise.

- (1) Each participant in the innovative program has the potential for experiencing a uniquely individualized "treatment."
- (2) The battery of diagnostic instruments or measuring procedures must have the potential to describe learner performance within highly individualized experiences.
- (3) The range of individual experiences must be so completely described so as to allow the identification of meaningful comparative programs.

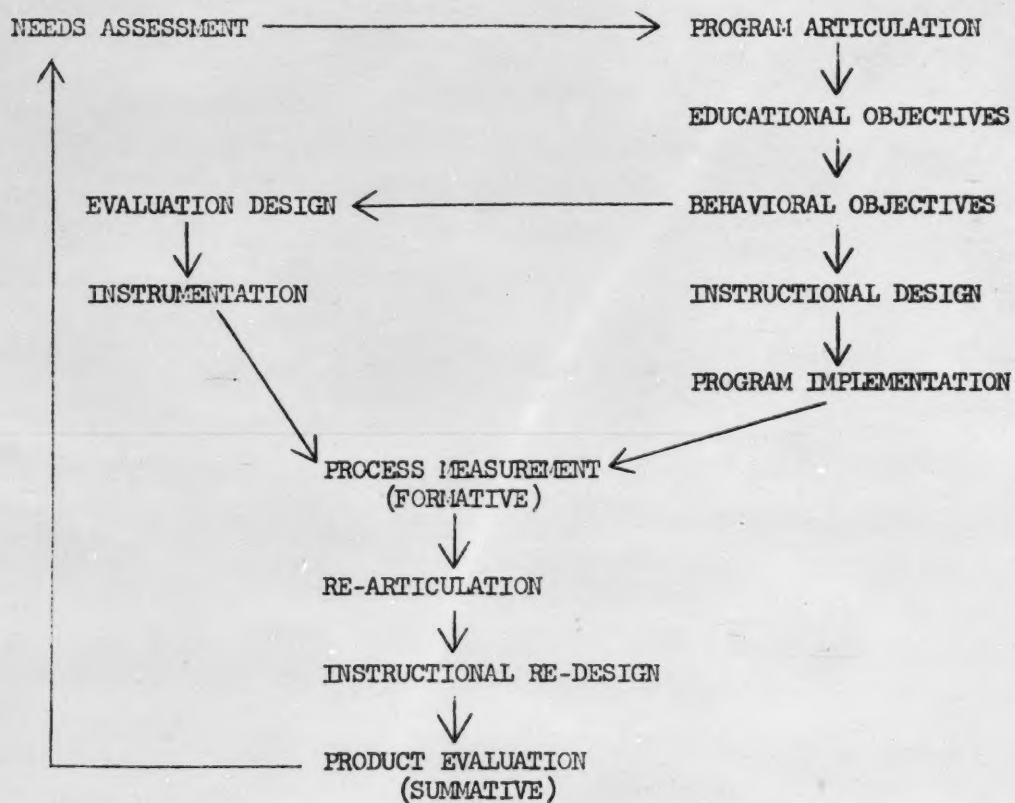
In essence, these three issues suggest that the evaluation of innovative programs is primarily an issue of definition. If one can define anticipated products, instruments can be developed or identified, and appropriate comparison groups can also be identified. However,

until products are clearly defined and operationalized, evaluation efforts are destined to yield meaningless and often negative findings. "If measurement is well articulated within a program, the program is likely to look effective" (Stodolsky, 1973, 7), thus, product measures should be closely tailored to program processes.

However, assuming that products or outcomes are well articulated, the definition of treatment in a uniquely individualized environment requires that individualized programs cannot be evaluated as group effects. If each participant has a unique experience, each participant's development or change or progress should be compared to either a specifically matched criterion or to a baseline rather than averaging products as if treatment intervention had been a shared experience, i.e. using the treatment group as the unit of analysis.

We have now returned to the issue critical to all non-experimental evaluation studies of the identification of appropriate comparison groups. The results of the case study reported here suggest that external comparisons do not yield meaningful evaluation of an innovative program. The results do, however, suggest that if one examines the stimuli for the program origination, one can re-examine the stimuli to determine if the program reduced the original needs. The stimuli may include skill acquisitions, skill deficit remediation, lack of instructional clarity, learner affective characteristics, and so forth. The "effectiveness" of the program would then be determined in light of the reductions of these stimuli. Thus, one begins to discern a drive-reduction evaluation model. And, if one's goals for program implementation is to reduce drives or needs, it becomes apparent that these stimuli must be articulated, treated, and re-articulated.

Following is an evaluation model that incorporates these ideas:



This model would permit the inclusion of cost-effectiveness data and would retain emphasis on program monitoring which appears to be essential to valid evaluation of innovative programs. It also allows flexibility in the choice of criteria for program "effectiveness." However, if there is a commitment to change and a commitment to enrichment as in the New Trier case study presented earlier, the issue of "effectiveness" must be determined internal to the program. It is this thrust that emphasizes the area of "Needs Assessment." And, if educators are committed to serving their constituency, program effectiveness can be determined only by the "fit" between program outcomes and expressed needs.

Conclusion

The purpose of this paper was to present a traditional-derived evaluation model for secondary education, the results of the application of that model, and the implications of the application for a "better" evaluation model. The first two parts of this stated purpose have been achieved. The third and critical part is only partially addressed. Whether this new and "better" evaluation model would really have yielded more insightful data about the New Trier program is an unanswered question. Whether this model will work effectively for other innovative programs is also an unanswered question. However, if one only gains insight into the problems inherent in the evaluation of innovative educational programs, something has been gained by the presentation of this paper. Clearly,

evaluation has begun to and will continue to examine the process aspects of program evaluation whether innovative or traditional. Clearly, unintended outcomes will continue to be an important part of evaluation studies. What is not clear is whether the financial constraints of Local Education Agencies (LEAs) will tolerate the added cost of useful evaluation studies. It would, indeed, be disastrous if evaluation was excluded from instructional development as a direct result of cost. However, the realities of the seventies may be that the only way to insure that evaluation becomes internalized in the educational planning mechanism is to require that teachers as well as administrators possess evaluative skills. The reliance upon outside "experts" to evaluate local educational programs may soon become an absurd idea.

1

Traditional in this sense means a common structure of the class-room learning: All children are taught at the same rate and are taught the same things. This learning environment earned the label of "structured-curriculum."

2

Learning Styles refers to preferences for personal interaction patterns while learning and not to modalities of information processing.

3

Critical thinking and problem-solving were represented by the Cornell Test of Critical Thinking rather than by two separate instruments. This occurred because adequate instruments for problem-solving skills were not available to the Community staff.

4

Although questionnaires are potentially insightful sources of data in descriptive evaluation studies, the one designed for this study failed to indicate change with respect to any of the eight affective behavioral objectives. Whether this was a function of "bad" questions, student apathy or, in fact, no change has not been determined. However, throughout all the questionnaires, whether from Community participants or from students in the parent school, a general dissatisfaction with "school" as an institution was apparent. And, although it is popular to reject the idea of institutionalized learning, it may be that the negative responses to school justify experimentation in alternative forms of education. Thus, this dissatisfaction may be one of the needs to be assessed.

5

Through initial interviews the Community faculty were able to identify those Community students who would be graduating in January.

6

The designation of comparison groups refers not to the status of the participating students except with respect to grade level, but to the status of the Community group with which the comparison group is paired.

7

The Preference for Structure Test (the shortened version of the Inventory of Beliefs) was scored so that higher scores indicate lack of preference.

References

- Bloom, B.S. Some theoretical issues relating to educational evaluation. In R.W. Tyler (Ed.), Educational evaluation: new roles, new means. Chicago: University of Chicago Press, 1969.
- Bloom, B.S. (Ed.). Taxonomy of educational objectives, handbook I: cognitive domain. New York: David McKay Co., Inc., 1956.
- Bock, R.D. Multivariate statistical methods in behavioral research. New York: McGraw-Hill, 1975.
- Buros, O.K. (Ed.). Mental measurement yearbook. Highland Park, New Jersey: Gryphon Press, 1971.
- Campbell, D.T. and Stanley, J.C. Experimental and quasi-experimental designs for research on teaching. In N.L. Gage (Ed.), Handbook of research on teaching. Chicago: Rand McNally and Co., 1963, 171-246.
- Cronbach, L.J. Essentials of psychological testing. New York: Harper and Row, Publishers, 1971.
- Educational evaluation decision making. Itasca, Illinois: F.E. Peacock Publishers, Inc., 1971.
- Karlson, A. and Stodolsky, S.S. Predicting outcomes from observations of child behavior in classrooms. American Educational Research Journal, 1973.
- Krathwohl, D.R.; Bloom, B.S.; and Masia, B.B. (Eds.). Taxonomy of educational objectives, handbook II: affective domain. New York: David McKay Co., Inc., 1964.
- Light, R.J. and Smith, P. Choosing a future: strategies for designing and evaluating new strategies. Harvard Educational Review, 1970.
- Lindvall, M. and Cox, M. Evaluating individualized instruction. In R.W. Tyler (Ed.), Educational evaluation: new roles, new means. Chicago: University of Chicago Press, 1969.
- Popham, W.J. (Ed.). Evaluation in education. Berkeley, California: McCutchan Publishing Corporation, 1975.
- Provus, M. Evaluation of ongoing programs in public school systems. In R.W. Tyler (Ed.), Educational evaluation; new roles, new means. Chicago: University of Chicago Press, 1969.
- Provus, M. The grand experiment. Berkeley, California: McCutchan Publishing Corporation, 1975.

Scriven, M. The methodology of evaluation. AERA Monograph Series on Curriculum Evaluation, 1971, 1.

Stodolsky, S.S. Defining treatment and outcomes in early childhood education. In B. Turner (Ed.) Education at Chicago. Chicago: Department of Education, 1973, spring.

Tyler, R.W. Basic principles of curriculum and instruction. Chicago: University of Chicago Press, 1971.

Appendix A

Description of Participants

Groups	Boys		Girls	
	11 grade	12 grade	11 grade	12 grade
I	7	12	11	21
II	0	0	0	3
III	2	0	0	3
IV	3	0	6	7
V	0	1	0	6
VI	0	2	1	0
Totals	12	16	18	40

Appendix B

Likert-type Scale

Directions: Following is a series of statements. You should read each statement carefully and then respond by circling one response choice. The response choices range from Strongly Agree (SA) to Agree (A) to Disagree (D) to Strongly Disagree (SD). This is not a time test. You will have enough time to finish all the items. Please answer all of the items with only one response. Please read each item thoroughly before choosing a response.

- | | | | | |
|--|----|---|---|----|
| 1. The goals of education should be determined by student need. | SA | A | D | SD |
| 2. People cannot fairly criticize another's work. | SA | A | D | SD |
| 3. Teachers understand students. | SA | A | D | SD |
| 4. No subject is more important than the personalities of the students. | SA | A | D | SD |
| 5. Evaluation of a student's work demonstrates what he has learned. | SA | A | D | SD |
| 6. Students shouldn't be told when and what to study. | SA | A | D | SD |
| 7. Learning is not experimental. | SA | A | D | SD |
| 8. Competition in school is wrong. | SA | A | D | SD |
| 9. Tests make people feel uncomfortable. | SA | A | D | SD |
| 10. Teachers encourage students to study and criticize their own work. | SA | A | D | SD |
| 11. Student's academic performance has nothing to do with whether teachers like them or not. | SA | A | D | SD |
| 12. What is taught in school is relevant to the adult world. | SA | A | D | SD |
| 13. Teachers help students by structuring classes. | SA | A | D | SD |
| 14. Evaluation doesn't show which students have learned the most. | SA | A | D | SD |
| 15. Competition motivates hard work. | SA | A | D | SD |

16.	Teachers don't care what students think.	SA	A	D	SD
17.	School is structured around irrelevant subjects.	SA	A	D	SD
18.	The heart of education is the process of learning, not the facts.	SA	A	D	SD
19.	Teachers like good students better than average students.	SA	A	D	SD
20.	People should compete with others.	SA	A	D	SD
21.	Alternative education gives the individual student independence.	SA	A	D	SD
22.	Having specified periods for different subjects interrupts the learning process.	SA	A	D	SD
23.	Education and educational institutions must be sources of social change.	SA	A	D	SD
24.	The goals of education should be specified by the school administration and the teacher.	SA	A	D	SD
25.	In order to be a successful adult, one must do well in school.	SA	A	D	SD
26.	Students need supervision in their work.	SA	A	D	SD
27.	Having one's work evaluated keeps one from exploring new subjects of interest.	SA	A	D	SD
28.	Subject matter is always interesting no matter what the class composition is.	SA	A	D	SD
29.	Having specified periods for studying helps one learn.	SA	A	D	SD
30.	Evaluations have no relationship to success after school	SA	A	D	SD
31.	Education and educational institutions do not reflect social change.	SA	A	D	SD
32.	Students are not capable of criticizing their own work.	SA	A	D	SD

33. Students don't worry about being evaluated when planning their academic program.	SA	A	D	SD
34. Good students hang around together.	SA	A	D	SD
35. Learning is discovery from your own experience.	SA	A	D	SD
36. The true view of education is that of a student who needs direction and a teacher who exerts direction and control.	SA	A	D	SD
37. People should compete with themselves and not with others.	SA	A	D	SD
38. Learning is more than accumulating facts.	SA	A	D	SD
39. Taking tests is fun.	SA	A	D	SD
40. School should give the individual the opportunity to do what they want.	SA	A	D	SD
41. Non-traditional education is interesting.	SA	A	D	SD
42. Teachers provide alternatives for the students.	SA	A	D	SD
43. Other people's criticism of one's work is helpful.	SA	A	D	SD
44. Traditional education is dull and tedious.	SA	A	D	SD

Adjective Ranking

Directions: Please rank the following words in order from 1 to 18 according to your estimate about how well each word describes your experience in school. One way to approach this problem is to read the list thoroughly. Then select the words that most accurately represent your positive and negative feelings about school. Label these words "1" for the most accurate and "18" for the least accurate. Then place the remaining sixteen words from 2 to 17 according to your best estimate of their value on the positive to negative scale. Remember to use each number only once.

Pleasant _____

Stifling _____

Boring _____

Exciting _____

Challenging _____

Uninteresting _____

Irrelevant _____

Fun _____

Structured _____

Flexible _____

Useful _____

Distasteful _____

Pointless _____

Repressive _____

Unpleasant _____

Relevant _____

Enjoyable _____

Accepting _____

STUDENT QUESTIONNAIRE

Directions: We want to understand the way you feel about yourself and about school. Please use complete answers for each question. Do not answer "Yes" or "No" without explaining why you feel that way.

Name _____

Grade _____

1. Do you perceive yourself as a poor, below average, average, good, or superior student?

2. Is it important to you to do well in school?

3. Are your friends participating in the Community for Alternatives in Education?

4. List three of your most important interests.

5. Where is your favorite place to study?

6. How long do you usually spend studying each day?

7. Do you like to study along? with one friend? or with a group of friends?

8. Are you self-confident in academic matters?

9. Do you feel that teachers are always judging you?

10. When you do well does someone notice you?

11. Do you worry about grades?

12. What is your favorite part of school?

13. What kind of learning experiences have you had this part year that have nothing to do with home or school?

14. Do you use the city? suburban? and/or rural environment, and if so, how?

15. Do you usually wear a watch?

Appendix C

Descriptive Information About Initial Pre-Test Battery

Test Name	Number Tested	Mean	Standard Deviation	KR ₂₀
Stanford Reading Pt. A	169	38.36	9.77	.93
Stanford Reading Pt. B	170	20.30	6.07	.93
Stanford English Pt. A	170	11.15	3.12	.78
Stanford English Pt. B	169	16.73	4.15	.86
Stanford English Pt. C	169	11.20	3.75	.88
Stanford English Pt. D	168	4.90	1.34	.63
Stanford English Pt. E	169	9.10	2.95	.87
Stanford Mathematics	170	30.24	11.14	.85
Rotter	169	6.10	6.15	.94
Cornell Test of Critical Thinking	169	42.30	14.47	.96

Appendix D

Test Name	Number Tested	Mean	Standard Deviation	Split-Half
Tennessee Self- Concept	169	209	23.54	.92
Teacher-Made Likert-Type Scale	167	157	21.80	.54
Adjective Ranking	132	*NA	*NA	.32

Note: *NA refers to "Not Applicable" because of nominal or ordinal data.

Appendix E
Interscale Correlation Matrix of the Stanford
Reading and English Tests
N=114

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Reading Part A (1)	1.00						
Reading Part B (2)	.65	1.00					
English Part A (3)	.34	.21	1.00				
English Part B (4)	.27	.28	.63	1.00			
English Part C (5)	.31	.14 ^{NS}	.57	.61			
English Part D (6)	.14	.12	.41	.46	.21	1.00	
English Part E (7)	.19 ^{NS}	.07 ^{NS}	.35	.35	.30	.26	1.00

Note: All correlation coefficients are significant except as marked "NS."

Appendix F

Suggestions for Evaluation of the Community for Alternatives
in Education for the 1975-1976 Academic Year

1. Select a random sample from the parent school (at least 200 to insure an adequate set of complete data). This group will serve as the one comparison group for a comparative evaluation desired by the parent school.
2. Schedule the post-test session early in the spring to avoid the distractions of spring for both the Community and comparison group.
3. Select one or two subscales from the Tennessee self-concept Test to be used in the study rather than all of the subscales.
4. Locate observers from outside the New Trier High School system to make periodic although unscheduled visits to the Community and to compile a notebook of narrative records to be used in assessing unintended outcomes and for process studies.
5. Have the Community participants submit monthly summaries of their learning and social experiences to the program director.
6. Exclude general ability measures from further evaluations since no differences were reported in the 1974-75 study. Additional inclusion of such instruments would result in wasted time needed for more highly relevant testing.
7. Construct and pilot additional affective instruments for objectives theoretically covered by the Student Questionnaire.
8. Attempt to communicate with other similar programs in both higher education and in secondary programs for evaluation guidelines and suggestions.
9. Locate technical assistance for psychometric analysis of piloted instruments within the school system or in the neighboring community.
10. Assessment of the effects of the Community structure of participating faculty.
11. Assessment of the relationship between Community faculty and the faculty of the parent school.

Note: Educational evaluation is not only the assessment of student learning. Evaluation should include intrinsic as well as pay-off comparisons. It is for this reason that this author recommends an emphasis on non-traditional measurement approaches in continued evaluation of this Community.